# LLM による日本語生成におけるモデル固有表現パターンの分析

林 美佐 [1]　相澤 彰子 [1,2]
[1] 東京大学 理学部情報科学科　[2] 国立情報学研究所
{m-hayashi1105,aizawa}@nii.ac.jp

## 概要

本研究では，同一の日本語プロンプトに対して複数の大規模言語モデル（LLM）が生成した応答文を分析し，モデル固有の出力特性が存在するかを検証した。商用 API モデルおよびオープンパラメータモデルを対象とし，生成文の埋め込み表現を用いた生成元モデルの分類，テキスト類似度指標によるモデル間比較，および頻出フレーズの抽出を行った。実験の結果，日本語生成テキストにおいても生成元モデルを高精度に識別可能であり，回答構造や言い回しにおいてモデルごとに特徴的な表現が出現することが観察された。

## 1　はじめに

近年，大規模言語モデル（LLM）の発展により，自然言語による文章生成技術が広く利用されるようになっている。それに伴い，生成されたテキストが人間によるものか，あるいはどの言語モデルによって生成されたものかを識別する重要性が高まっている。人間と LLM 生成文の識別に関する研究は数多く行われている一方で [1, 2]，「どの LLM が生成したか」を識別する研究は相対的に少ない。

このような背景のもと，同一の入力に対する複数の大規模言語モデルの出力を分析する研究が行われ，生成文に含まれる意味的特徴および表層的特徴を手がかりとして，出力元モデルを高精度に識別できることが報告されている [3]。しかし，多くの既存研究は英語を対象としており，英語において報告された高い識別性能が，日本語生成テキストにおいても同様に得られるかについては，十分に検証されていない。[4] においても指摘されているように，日本語は英語とは文法構造や表現形式が大きく異なる。また，LLM の学習データの量や質にも言語間で差が存在するため，英語で得られた知見が日本語にもそのまま成立するとは限らないことが示唆されている。

本研究では，日本語における LLM の出力特性を明らかにすることを目的とする。具体的には，複数の LLM に同一の日本語プロンプトを与えて得られた生成文を入力とし，出力元モデルを推定する分類実験を行う。商用 API モデルおよびオープンパラメータモデルを対象とし，日本語生成テキストにおいてもモデル固有の出力特性が観測されるかを検証する。

## 2　実験手法概要

### 2.1　使用データセット

実験には，2 種類のインストラクションデータセットを用いた。

1 つ目は Ichikara [5] であり，日本語で構築されたインストラクションデータセットである。本研究では，質問文のみを抽出し，重複を除去した 3,985 件のプロンプトを使用した。

2 つ目は UltraChat [6] である。英語で構築されたデータセットであるため，ランダムに抽出した質問文を日本語へ翻訳して使用した。翻訳および品質評価には `gpt-4o-mini` を用い，原文と翻訳文の組に対して 100 点満点で評価を行い，95 点以上のもののみを採用した。最終的に Ichikara と同数の 3,985 件の日本語プロンプトを構築した。

### 2.2　使用したモデル

生成モデルとして，商用 API モデルおよびオープンパラメータモデルを用いた。使用したモデルの一覧を表 1 に示す。特に断りがない限り，商用 API モデルはデフォルト設定で使用し，オープンパラメータモデルは確率的サンプリングによる出力の揺らぎを抑え，モデル固有の出力特性を安定して観測するため，temperature を 0 に設定して生成を行った。

**表 1**　実験に使用したモデル一覧

商用 API モデル

| Model | Params | Model ID |
|---|---|---|
| GPT-5.2 | – | gpt-5.2-2025-12-11 |
| Claude Sonnet | – | claude-sonnet-4-5-20250929 |
| Gemini 3 Pro | – | gemini-3-pro-preview |

オープンパラメータモデル

| Model | Params | Model ID |
|---|---|---|
| LLaMA-3.1 | 8B | meta-llama/Meta-Llama-3.1-8B-Instruct |
| Gemma-2 | 9B | google/gemma-2-9b-it |
| Qwen2.5 | 7B | Qwen/Qwen2.5-7B-Instruct |
| Mistral-7B | 7B | mistralai/Mistral-7B-Instruct-v0.3 |
| LLM-jp-3.1 | 13B | llm-jp/llm-jp-3.1-13b-instruct4 [7] |
| Swallow-7B | 7B | tokyotech-llm/Swallow-7b-instruct-v0.1 [5] |

## 2.3　分類手法

複数の LLM に同一プロンプトを与えて得られた生成文を入力とし，生成元モデルをラベルとする多クラス分類問題を構成した。

生成文は埋め込みモデル Ruri [8] により文ベクトルへ変換し，最終層の隠れ状態に対する mean pooling を用いた表現を線形分類ヘッドに入力して予測を行った。

## 2.4　実行環境

分類実験は，NVIDIA A100（80GB）を搭載した GPU サーバ上で、accelerate launch により 2 GPU の分散学習として実行した。特に断りのない限り，バッチサイズ 8，エポック数 10，学習率 1e-5 とした。学習用・評価用サンプル数はそれぞれ 3,485 件および 500 件である。

# 3　実験結果

本節では，第 2 節で述べた実験手法に基づき，大規模言語モデル（LLM）ごとの日本語生成テキストの特徴を分析する。埋め込み表現を用いた分類実験，テキスト類似度指標による分析，および頻出フレーズの抽出結果を示す。

## 3.1　埋め込み分類によるモデル識別性能

生成文の埋め込み表現を入力とする多クラス分類により，日本語生成テキストから生成元モデルを識別できるかを評価した。

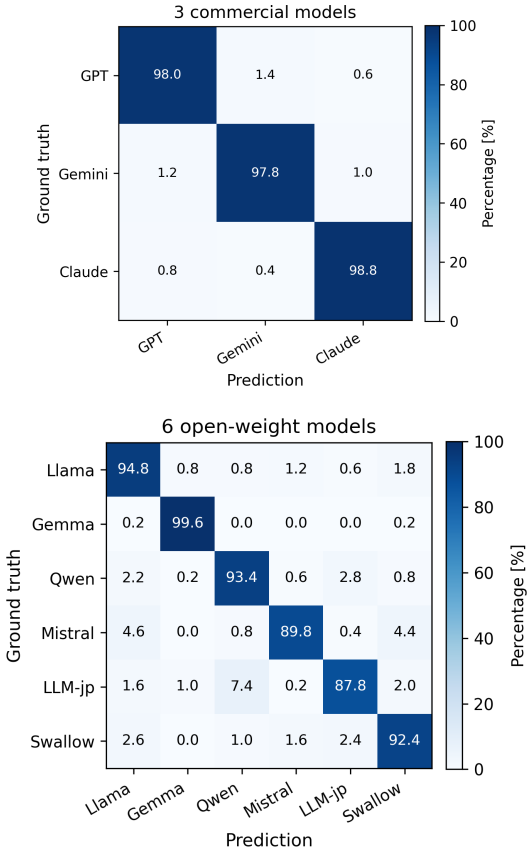


**図 1**　埋め込み分類によるモデル識別の混同行列．
上：商用 API モデル 3 種類を対象とした 3 クラス分類の結果．下：オープンパラメータモデル 6 種類を対象とした 6 クラス分類の結果．

### 3.1.1　モデル種別ごとの生成元分類結果

生成文の埋め込み表現を用いて，商用 API モデル 3 種類（表 1 上）およびオープンパラメータモデル 6 種類 (表 1 下）を対象に分類実験を行った。商用 API モデルの 3 クラス分類では分類精度は 98.2% であった。オープンパラメータモデルの 6 クラス分類においても，分類精度は 93.0% であった。

対応する正規化混同行列を図 1 に示す。商用 API モデル（上）では正解クラスへの予測確率が 98.2% と高く，誤分類は一部に限られている。オープンパラメータモデル（下）においても，6 クラス分類の設定下で 93.0%と高い識別性能が確認された。

### 3.1.2　データセット変更時の分類性能の比較

使用するデータセットの違いが分類性能に与える影響を調べるため，Ichikara および UltraChat を用いて同一条件で分類実験を行った。得られた分類精度を以下に示す。

- Ichikara（日本語）：93.0%

表 2　バージョン差に基づく分類結果

| モデル | バージョン | Acc.(%) |
|---|---|---|
| **ChatGPT** | gpt-5.1-2025-11-13<br>gpt-4o-2024-08-06<br>o3-2025-04-16<br>gpt-5.2-2025-12-11 | 93.6 |

表 3　推論設定の違いによる分類結果

| モデル | 推論設定 (thinking_level) | Acc.(%) |
|---|---|---|
| **Gemini** | high<br>low | 73.0 |

- UltraChat（英語）：99.3%
- UltraChat（日本語翻訳）：99.4%

## 3.2　モデルファミリー内での識別可能性

同一系列に属する LLM 間において，生成文に基づく識別が可能であるかを検証した。商用 API モデルのバージョン差，推論設定の違い，およびオープンパラメータモデルのサイズ差に着目した分類実験を行った。

### 3.2.1　バージョン差に基づく分類結果

GPT シリーズ 4 種類を対象として 4 クラス分類を行った。使用したモデルおよび分類精度を表 2 に示す。分類精度は 93.6% であった。

### 3.2.2　推論設定の違いに基づく分類結果

Gemini モデルの thinking_level を high および low に設定した 2 種類を対象として 2 クラス分類を行った。対象モデルおよび分類精度を表 3 に示す。

### 3.2.3　サイズ違いのモデルの分類結果

オープンパラメータモデルにおいて，モデルサイズの違いに基づく分類実験を行った。対象モデルおよび分類精度を表 4 に示す。いずれのモデル系列においても分類精度は 75% 以上であった。

## 3.3　テキスト類似度指標によるモデル間差異の検証

テキスト類似度指標を用いて，同一モデル内および異なるモデル間における生成出力の類似度を比較した。類似度指標として，ROUGE-1 [9]，ROUGE-L [9]，および BERTScore [10] の 3 種類を用いた。

表 4　モデルサイズの違いによる分類結果

| 系列 | モデル | Acc.(%) |
|---|---|---|
| **Llama** | Llama-3.1-8B-Instruct<br>Llama-3.1-70B-Instruct | 96.9 |
| **Qwen** | Qwen2.5-0.5B-Instruct<br>Qwen2.5-3B-Instruct<br>Qwen2.5-7B-Instruct<br>Qwen2.5-14B-Instruct<br>Qwen2.5-32B-Instruct | 88.0 |
| **LLM-jp** | llm-jp-3.1-1.8b-instruct4<br>llm-jp-3.1-13b-instruct4 | 79.9 |
| **Swallow** | Swallow-7b-instruct-v0.1<br>Swallow-13b-instruct-v0.1 | 75.5 |

### 3.3.1　実験設定

同一の質問に対する応答を，3 種類の商用 API モデルからそれぞれ 2 回ずつ生成し，計 6 個の出力を得た。得られた全ての出力ペアについて類似度を計算し，モデル内およびモデル間で比較した。出力には事前処理として MeCab による形態素解析を適用した。

### 3.3.2　結果

各指標に基づく類似度行列を図**??**に示す。いずれの指標においても，同一モデルから生成された出力同士のペアは，異なるモデル間の出力ペアと比較して高い類似度を示した。

ROUGE-1 および ROUGE-L では，同一モデル内の出力ペアが近い値を示し，BERTScore においても同様の傾向が確認された。

## 3.4　頻出フレーズの抽出

word n-gram に基づき，各モデルにおける頻出フレーズを抽出した。出力には，前節と同様に MeCab による形態素解析を適用した。頻出フレーズは，以下のスコアに基づいてランキング化した。

$$\text{score} = \frac{\text{freq}_{\text{class}} + \alpha}{\text{freq}_{\text{other}} + \alpha} \times 2^{n} \tag{1}$$

ここで，$\text{freq}_{\text{class}}$ は対象モデルにおいて，当該 n-gram を含む生成応答（プロンプトに対する回答）の数を表し，同一応答内で複数回出現した場合でも 1 として数える。$\text{freq}_{\text{other}}$ はそれ以外のモデルにおける同様の出現数である。$\alpha$ は平滑化項であり，$n$ は n-gram の長さを表す。

本スコアに基づいて得られた上位の n-gram のうち，重複を除去した代表的なフレーズを抽出した。これらのフレーズは，各モデルにおいて高い頻度で

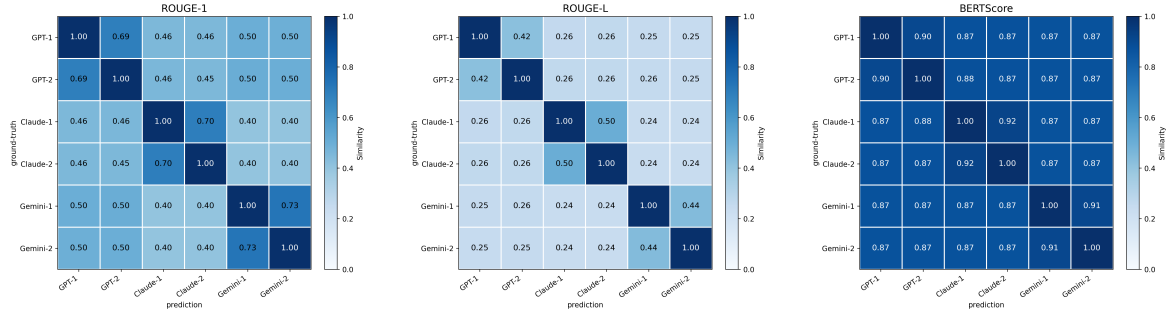


**図 2**　テキスト類似度指標に基づく類似度行列．左から順に，ROUGE-1，ROUGE-L，BERTScore による結果を示す．

用いられる表現傾向を示す例であり，抽出された全ての n-gram を列挙するものではない。代表的な例は付録（Appendix A）に示す。

抽出結果を概観すると，モデルごとに回答の構造や言い回しに違いが見られた。例えば，ChatGPT では丁寧な依頼表現を含む導入が頻出する一方，Gemini では結論提示や番号付き構造を伴う表現が多く観測された。また，LLM-jp 系モデルでは，「ステップバイステップで説明します。」といった手順的な説明を明示する導入表現が特徴的に現れていた。

## 4　分析

本研究では，日本語生成を対象として，複数の手法により大規模言語モデル（LLM）間の差異を検証した。

埋め込み表現を用いた分類実験において，商用 API モデルおよびオープンパラメータモデルの双方で高い識別性能が得られたことから，日本語生成テキストにおいてもモデル固有の出力特性が一貫して存在することが示唆される。

データセットを変更しても高い分類精度が維持されたことは，出力特性が特定のデータセットに依存せず，モデル自体に由来する可能性を示している。また，同一モデルファミリー内では，バージョン差やモデルサイズは識別可能であった一方，推論設定による差異は相対的に小さかった。

さらに，テキスト類似度指標に基づく分析から，同一モデル内で生成された出力同士は，異なるモデル間の出力と比較して，語彙選択や回答構造の点で高い一貫性を示すことが確認された。一方，頻出フレーズの分析に着目すると，ChatGPT では丁寧な依頼表現を伴う導入，Gemini では結論提示や番号付き構造を用いた説明，LLM-jp 系モデルでは手順的な説明を明示する導入表現など，モデルごとに特徴的な表現傾向が観測された。これらの結果は，埋め込み表現に依存しない表層的な言語使用のレベルにおいても，モデル固有の出力特性が現れていることを示している。

## 5　おわりに

本研究では，日本語生成テキストを対象として，大規模言語モデル（LLM）ごとに一貫した出力特性が存在することを実証的に示した。これは，英語を主対象として報告されてきたモデル固有性に関する既存研究の知見を，日本語へ拡張する結果である。

近年，LLM は教育・研究・行政・創作などで利用されており，生成されたテキストの出所を理解・検証するための技術的基盤の重要性が高まっている。本研究で示したモデル固有の出力特性は，生成元モデルの推定や出力分析の可能性を示すものである。

一方で，生成元モデルが推定可能であるという事実は，モデルの出自が未知であることを前提としてきた既存のモデル評価の枠組みに対し，再検討の余地を与える。今後は，出力特性の識別可能性を踏まえ，より公平かつ妥当な LLM 評価手法を検討する必要がある。

## 謝辞

本研究の一部は JSPS 科研費 24K03231 の助成を受けて行いました。

## 参考文献

[1] Tony Berber Sardinha. Ai-generated vs human-authored texts: A multidimensional comparison. **Applied Corpus Linguistics**, Vol. 4, No. 1, p. 100083, 2024.

[2] Wataru Zaitsu and Mingzhe Jin. Distinguishing chatgpt(-3.5, -4)-generated and human-written papers through japanese stylometric analysis. **PLOS ONE**, Vol. 18, No. 8, pp. 1–12, 08 2023.

[3] Mingjie Sun, Yida Yin, Zhiqiu Xu, J. Zico Kolter, and Zhuang Liu. Idiosyncrasies in large language models, 2025.

[4] Myung Hye Yoo, Joungmin Kim, and Sanghoun Song. Multilingual capabilities of gpt: A study of structural ambiguity. **PLOS One**, Vol. 20, , 2025.

[5] 井之上直也, 安藤まや, 後藤美知子, 関根聡, 中山功太, 宮尾祐介. 日本語を対象とした llm の大規模人手評価. 言語処理学会第 31 回年次大会発表論文集, 2025.

[6] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations, 2023.

[7] LLM-jp, :, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, Hiroki Kanezashi, Hiroshi Kataoka, Satoru Katsumata, Daisuke Kawahara, Seiya Kawano, Atsushi Keyaki, Keisuke Kiryu, Hirokazu Kiyomaru, Takashi Kodama, Takahiro Kubo, Yohei Kuga, Ryoma Kumon, Shuhei Kurita, Sadao Kurohashi, Conglong Li, Taiki Maekawa, Hiroshi Matsuda, Yusuke Miyao, Kentaro Mizuki, Sakae Mizuki, Yugo Murawaki, Akim Mousterou, Ryo Nakamura, Taishi Nakamura, Kouta Nakayama, Tomoka Nakazato, Takuro Niitsuma, Jiro Nishitoba, Yusuke Oda, Hayato Ogawa, Takumi Okamoto, Naoaki Okazaki, Yohei Oseki, Shintaro Ozaki, Koki Ryu, Rafal Rzepka, Keisuke Sakaguchi, Shota Sasaki, Satoshi Sekine, Kohei Suda, Saku Sugawara, Issa Sugiura, Hiroaki Sugiyama, Hisami Suzuki, Jun Suzuki, Toyotaro Suzumura, Kensuke Tachibana, Yu Takagi, Kyosuke Takami, Koichi Takeda, Masashi Takeshita, Masahiro Tanaka, Kenjiro Taura, Arseny Tolmachev, Nobuhiro Ueda, Zhen Wan, Shuntaro Yada, Sakiko Yahata, Yuya Yamamoto, Yusuke Yamauchi, Hitomi Yanaka, Rio Yokota, and Koichiro Yoshino. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms, 2024.

[8] Hayato Tsukagoshi and Ryohei Sasano. Ruri: Japanese general text embeddings, 2024.

[9] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

# A　頻出フレーズの一覧

**表 5**　モデルごとに抽出された代表的な頻出フレーズ

| モデル | 頻出フレーズ例 |
|---|---|
| ChatGPT | もしよければ，次を教えてください。 |
| | 差し支えなければ，次を教えてください。 |
| | を教えてください。状況に合わせて「 |
| | を教えてください。条件に合わせて「 |
| Gemini | 解説します。— 1 . |
| | 以下の通りです。 1 . |
| | まとめました。 1 . |
| | 結論から申し上げますと， |
| Claude | ていただければ，より具体的なアドバイスができます |
| | 方法を試してみてください。 |
| | 具体的に知りたいことはありますか？ |
| | 主な理由は以下の通りです： |
| Llama | 方法を試してみることができます。 |
| | 以下のようになっています。 |
| | 以下の点を考慮する必要 |
| Gemma | いくつかご紹介します！ |
| | いくつか質問させてください。 |
| | これらの情報があれば，あなたにぴったりの |
| | 予算はどのくらいですか？ |
| Qwen | いくつかのアドバイスを提供します：1 |
| | 以下にいくつかの提案をします：1 |
| | いくつか挙げてみます：1 |
| Mistral | の方法を考慮してみてください。1 |
| | を選択することをお勧めします。 |
| | の手順を参考にしてください。1 |
| LLM-jp | ステップバイステップで説明します。 |
| | に相談することも一つの方法です。 |
| | について，一般的なことを述べます。 |
| | 以下のポイントを考慮すると良いでしょう。 |
| Swallow | に応じて選ぶことをお勧めします。 |
| | 医療従事者に相談することをお勧めします。 |
| | 予算に応じて選ぶことをお勧めします。 |
| | に基づいて選ぶことをお勧めします。 |